

Master in Artificial Intelligence



Data Collection & Preprocessing I





Purpose

The purpose of the section is to help you learn how to collect and preprocess data to become a Successful Artificial Intelligence (AI) Engineer

At the end of this lecture, you will learn the following

- **How to gather relevant data from various sources, ensure its quality, and preprocess it to make it suitable for analysis and modeling**



How to collect and preprocess data

Gather
relevant
data

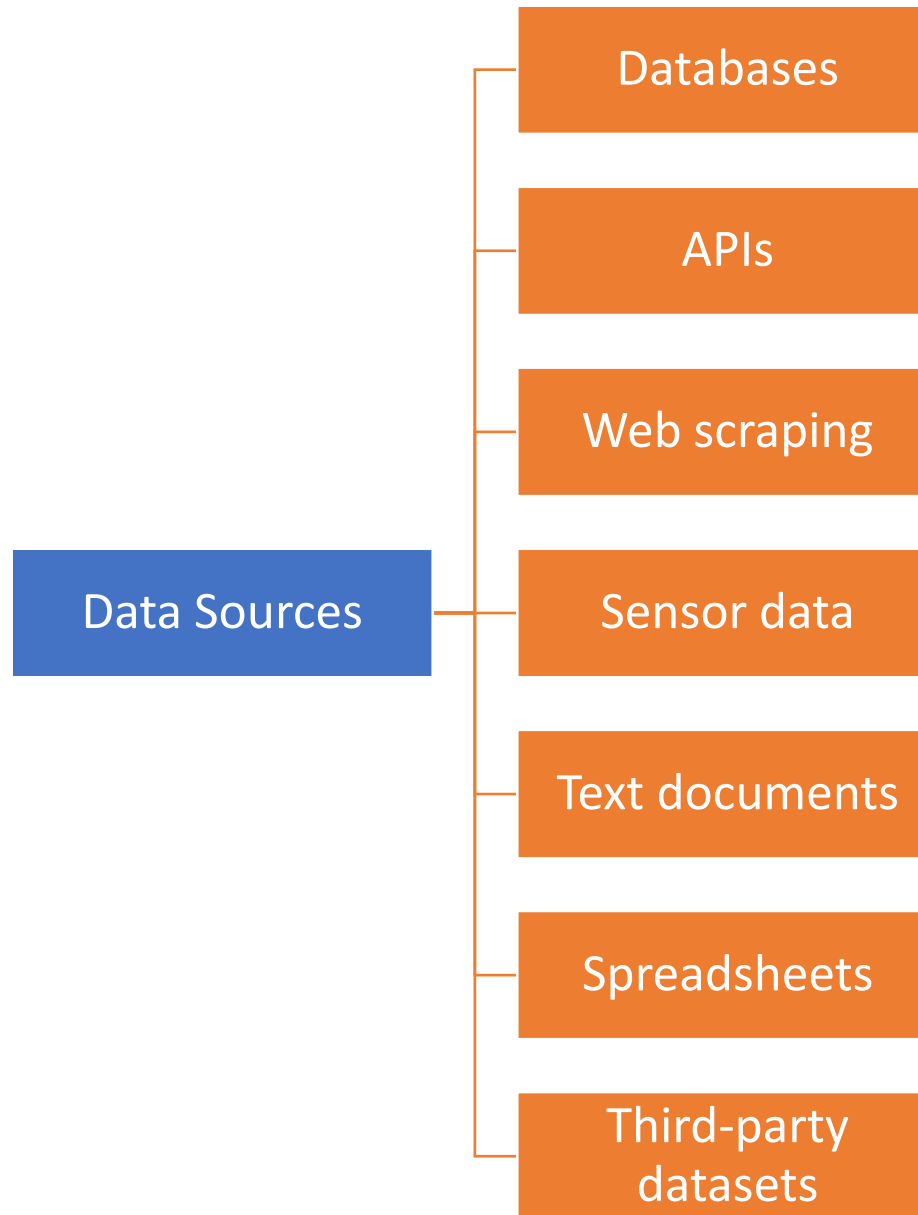
Ensure its
quality

Preprocess
it

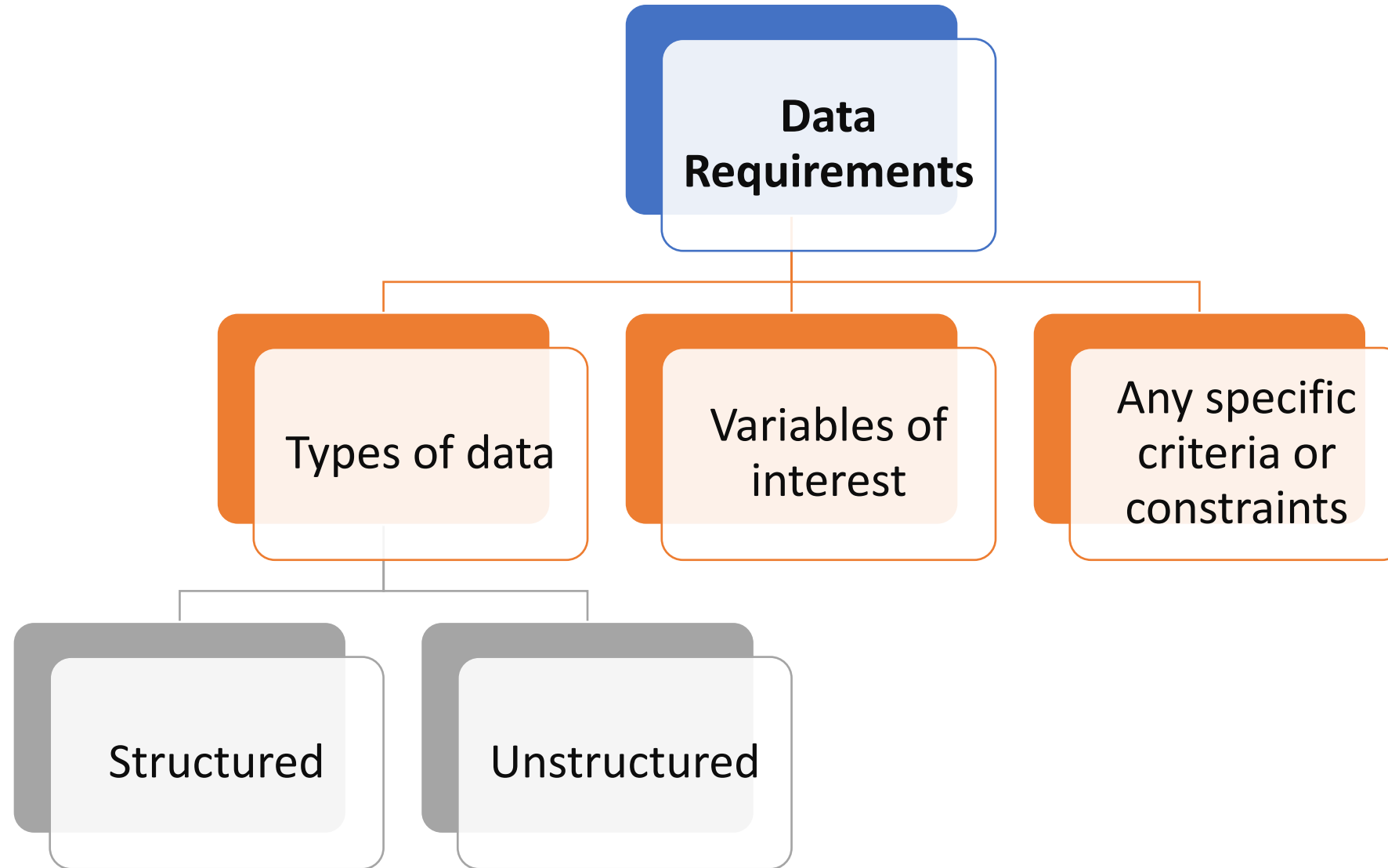
Make it
suitable for
analysis and
modeling.



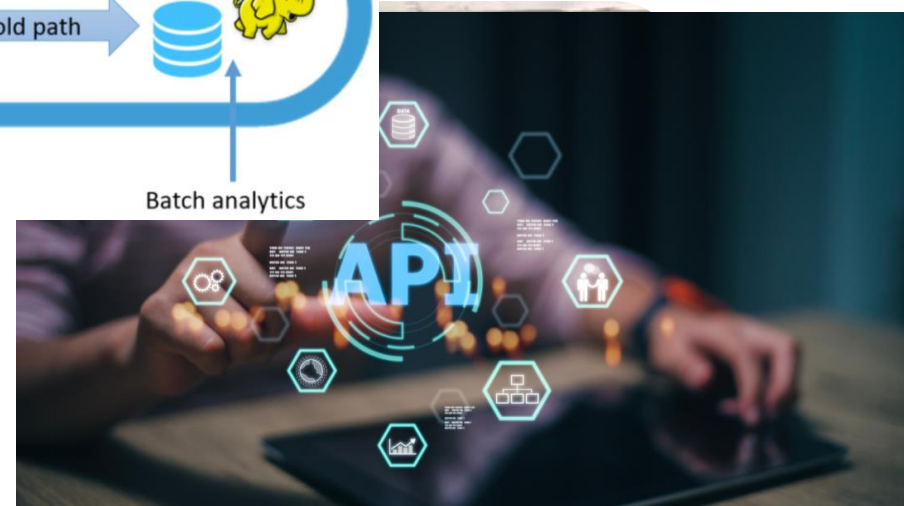
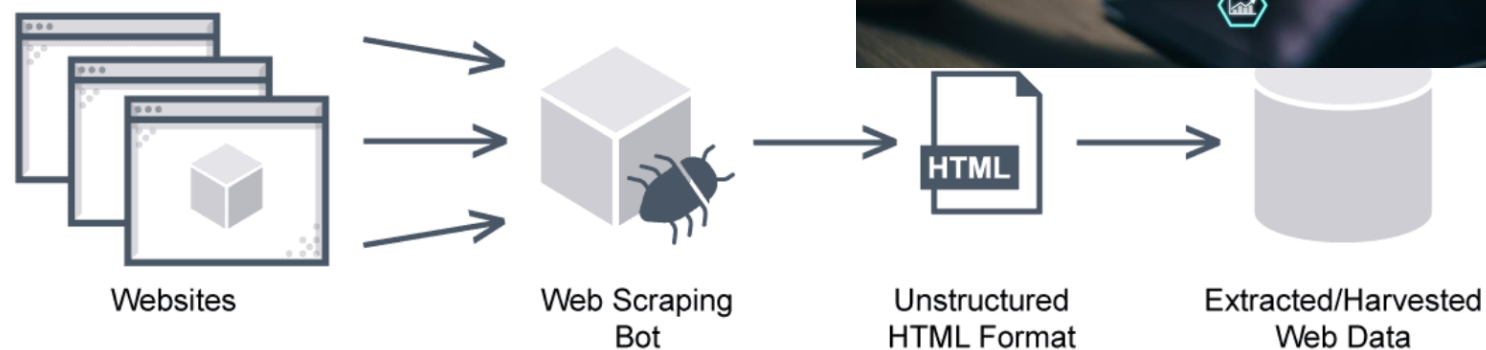
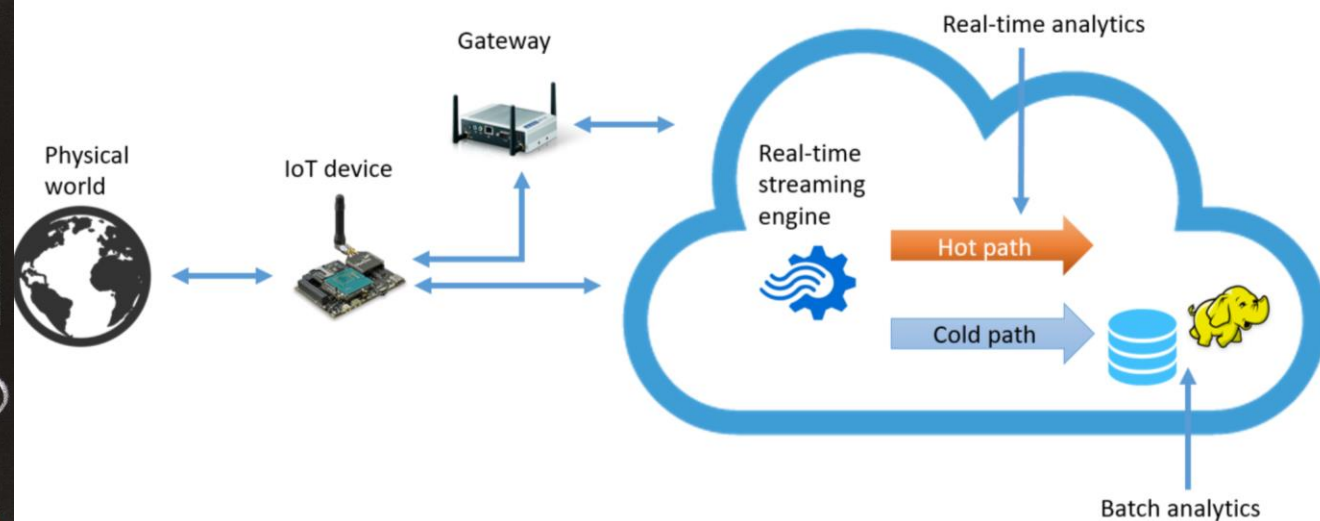
Identify Data Sources



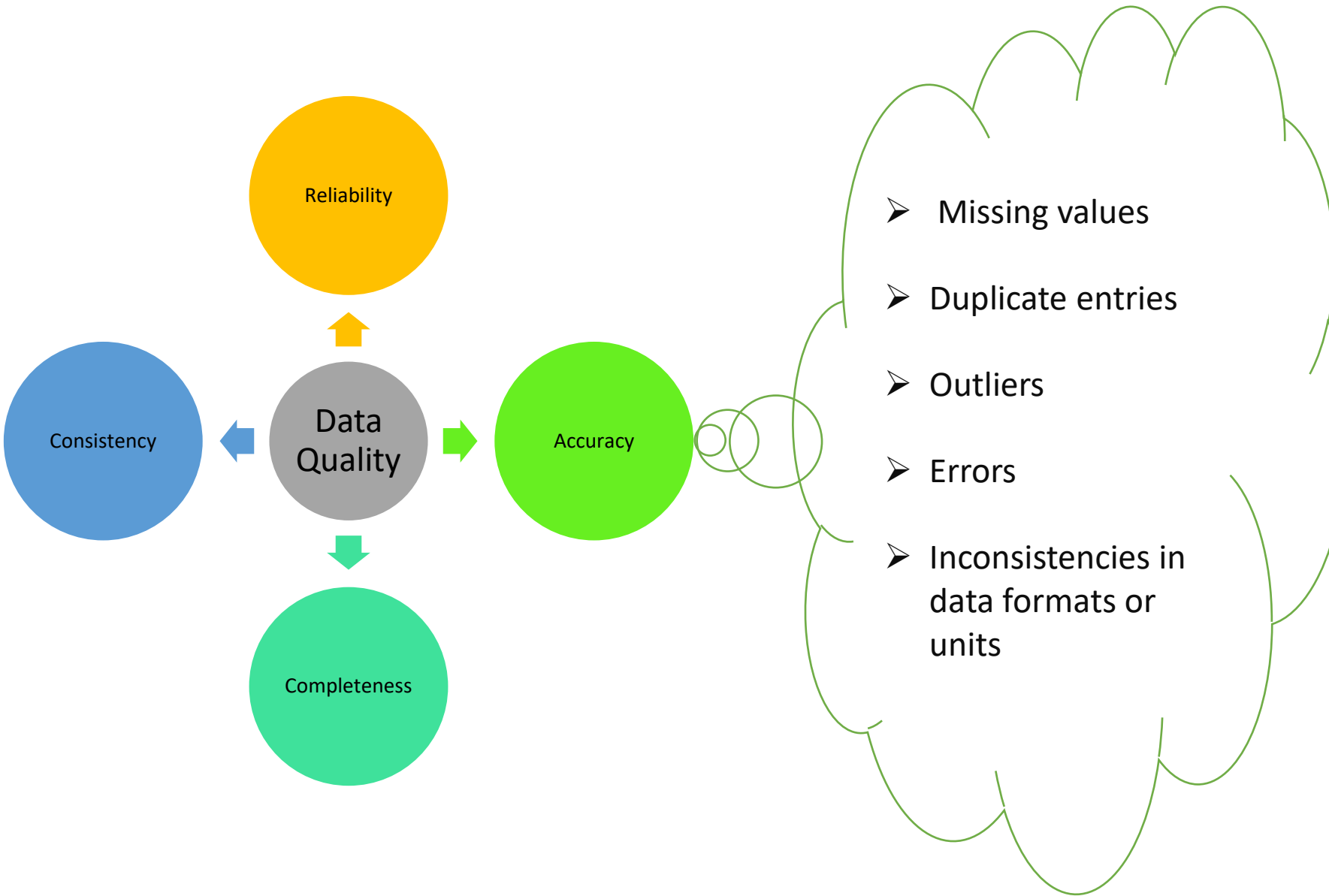
Define Data Requirements



Collect Data



Assess Data Quality



Clean Data

- Missing values
- Duplicate entries
- Outliers
- Errors
- Inconsistencies in data formats or units



Handling Missing Values

Statistical
methods

Mean

Median

Mode

Deleting
missing
data

Rows

Columns

Using
advanced
imputation
techniques

K-nearest
neighbors

Regression
imputation



Mean, Median, Mode

7, 3, 4, 1, 7, 6

Sum of numbers divided
by the total numbers

$$\text{Mean} = (7+3+4+1+7+6)/6 \\ = 28/6 = 4.66$$

7, 3, 4, 1, 7, 6

Arrange in order and
pick the middle value

1, 3, 4, 6, 7, 7

$$\text{Median} = (4+6)/2 = 5$$

Mode

7, 3, 4, 1, 7, 6

Most common number

7, 3, 4, 1, 7, 6

$$\text{Mode} = 7$$

Range

7, 3, 4, 1, 7, 6

Difference between
highest and lowest

$$\text{Range} = 7 - 1 = 6$$

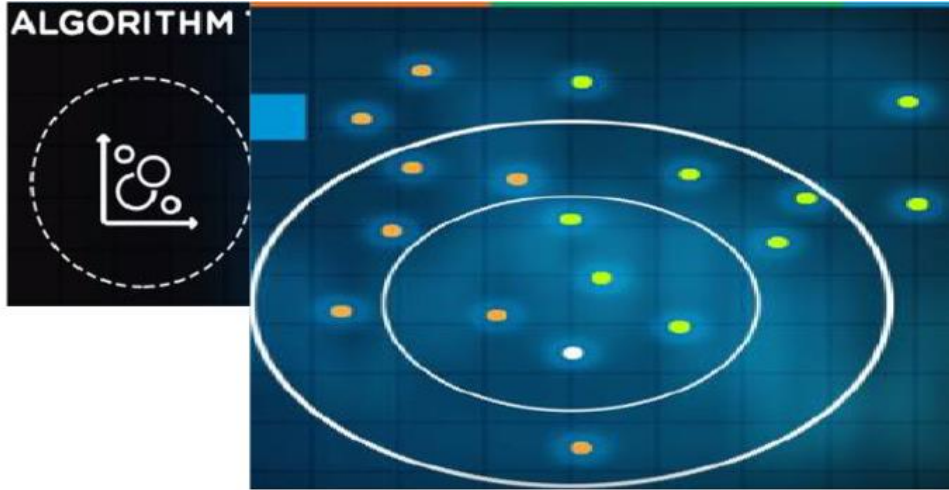


@gofodu

More like



K-nearest neighbors



K-NN

K-Nearest Neighbors (KNN) is a simple and versatile machine learning algorithm used for both **classification and regression tasks**.

It's based on the **principle of similarity**, where the predicted label or value of a new data point is determined by the labels or values of its k nearest neighbors in the training dataset.



Regression Imputation

Regression Imputation (Buck, 1960)

Replaces missing values with predicted scores from a regression equation by using information from the complete variables.

Advantages

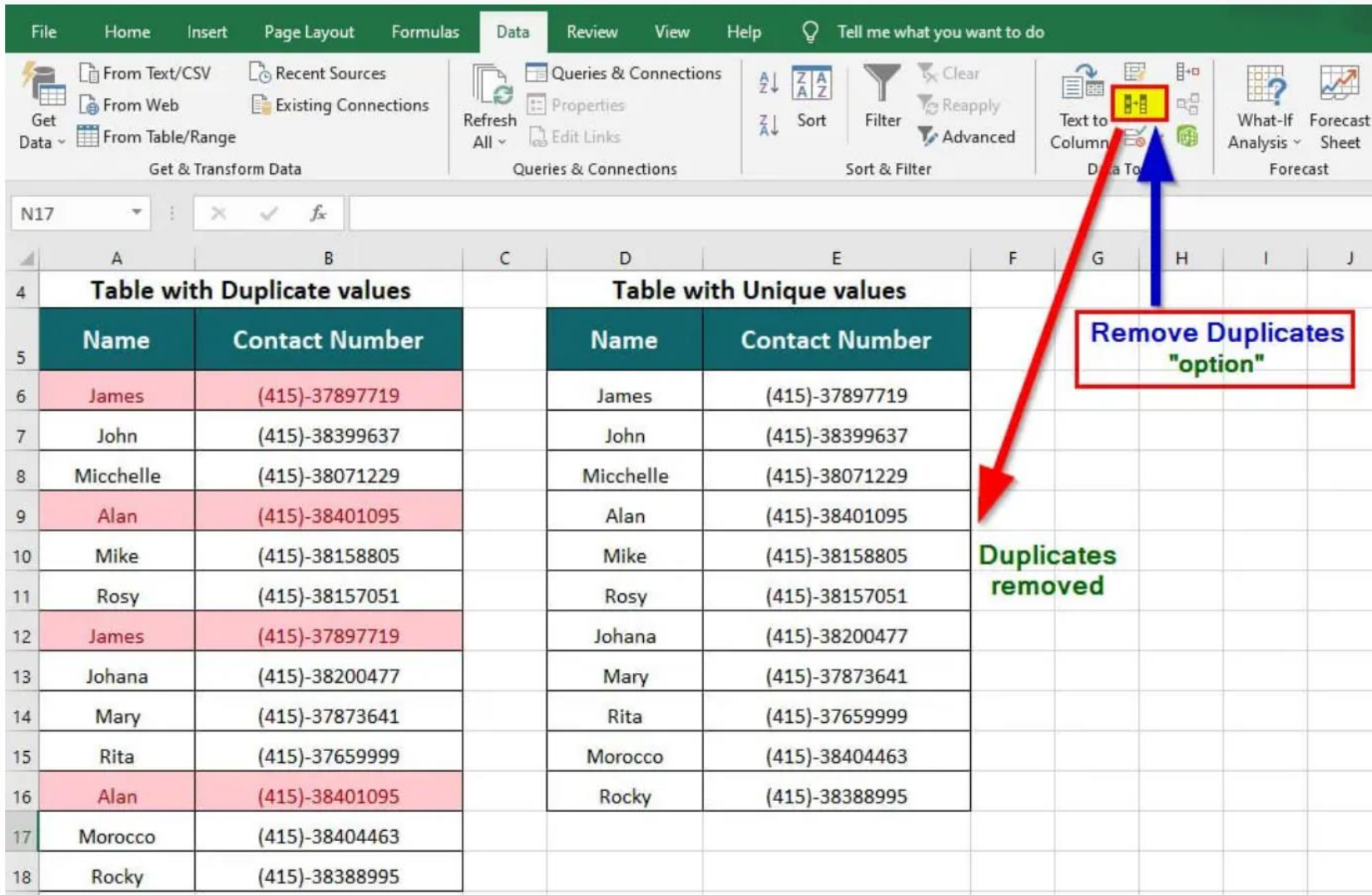
- It generates a complete data set.
- Variables tend to be correlated

Disadvantages

- Inputs data with perfectly correlated scores
- Over estimate correlation



Removing duplicates



The screenshot shows the Microsoft Excel interface with the 'Data' tab selected. The 'Remove Duplicates' button is highlighted in the 'Data Tools' group. A red arrow points from this button to the 'Table with Duplicate values' table, and a blue arrow points from the button to the 'Remove Duplicates "option"' text box.

Table with Duplicate values		Table with Unique values	
Name	Contact Number	Name	Contact Number
James	(415)-37897719	James	(415)-37897719
John	(415)-38399637	John	(415)-38399637
Micchelle	(415)-38071229	Micchelle	(415)-38071229
Alan	(415)-38401095	Alan	(415)-38401095
Mike	(415)-38158805	Mike	(415)-38158805
Rosy	(415)-38157051	Rosy	(415)-38157051
James	(415)-37897719	Johana	(415)-38200477
Johana	(415)-38200477	Mary	(415)-37873641
Mary	(415)-37873641	Rita	(415)-37659999
Rita	(415)-37659999	Morocco	(415)-38404463
Alan	(415)-38401095	Rocky	(415)-38388995
Morocco	(415)-38404463		
Rocky	(415)-38388995		

Remove Duplicates "option"

Duplicates removed

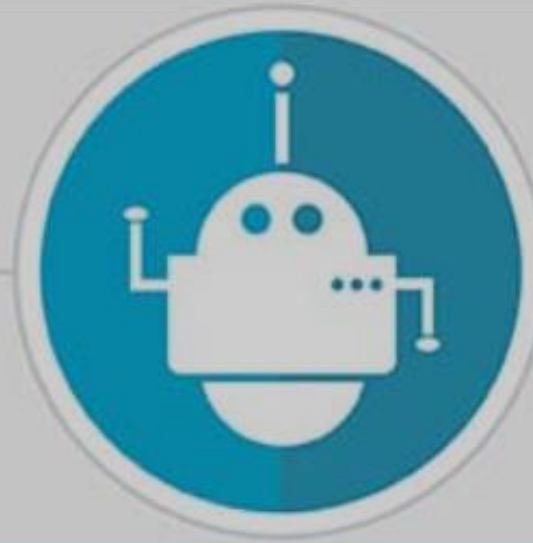


Correcting errors



Programmable Bots

Fetch data for processing, moving files & folders to data storages for easy access, accuracy & minimal data loss



Trained Bots

Data capture, calculations according to logical "if/then" rules, efficient copy paste, merging data from multiple sources.



ML & AI

Connects XML files to specific legacy systems to automatically populate the data in the correct fields.



What is next?

Standardizing data formats

	Values in Different Formats	Standardized Data
Dates	Dates in international document: <ul style="list-style-type: none">• 12/31/2023• 21-12-2023• 2023/11/15	ISO 8601 Format: YYYY/MM/DD <ul style="list-style-type: none">• 2023/12/31• 2023/12/21• 2023/11/15
Measurement Units	Different weight formats: <ul style="list-style-type: none">• 150 pounds• 64 oz• 11 stone	Standard Format: KG <ul style="list-style-type: none">• 68.039 kg• 1.814 kg• 69.853 kg
Language Translation	Phrases from news articles in different languages	All content translated to English for consistency



Master in Artificial Intelligence



Data Collection & Preprocessing I

